CLAIMS

1. A method of generating a database of molecular fragment data, the molecular fragment data defining at least parts of a number of molecular structures, from a data set retained within a store, the data set containing predetermined molecular structure data defining a number of predetermined molecular structures, the method comprising:-

a) selecting first and second molecular structure data defining first and second molecular structures respectively from the data set;

b) comparing the selected first and second molecular structure data to determine molecular fragment data, wherein the molecular fragment data defines at least part of a molecular structure common to each of the first and second molecular structures;

c) storing the determined molecular fragment data in the data set; and,

d) repeatedly performing steps (a) to (c) wherein either of the first or the second molecular structure data is selected from either the predetermined molecular structure data or the molecular fragment data determined in step (b) such that the resultant data set comprises a database containing the molecular fragment data.

2. A method according to claim 1, wherein the molecular fragment data determined in step (b) defines the maximum molecular structure common to the first and second molecular structures.

3. A method according to claim 2, wherein the maximum common structure is the maximum common molecular substructure between the first and second molecular structures.

4. A method according to claim 2 or claim 3, wherein the maximum common structure is the maximum molecular structure of atoms bonded as a single connected entity.

5. A method according to any of the preceding claims, further comprising a step of comparing the molecular fragment data determined in step (b) with the molecular structure data already within the data set and subsequently storing the determined molecular fragment data only if molecular structure data defining an identical molecular structure is not already present within the data set.

6. A method according to claim 5, wherein the comparison between the determined molecular fragment data and the molecular structure data is based upon the molecular masses of the corresponding molecular fragment and the molecular structure respectively.

7. A method according to any of the preceding claims, wherein the determined molecular fragment data comprises data identifying the first or second molecular structure data which has been involved in any previous comparison steps which have resulted in the determination of that particular molecular fragment data.

8. A method according to claim 7, wherein the identifying data is retained as a parent list identifying the molecular structure data.

9. A method according to claims 7 or claim 8, and when dependent upon claim 5, wherein, if the determined molecular fragment as defined by the molecular fragment data, is found to be identical to a molecular structure as defined by corresponding molecular structure data already present within the data set, the identifying data is added to the molecular structure data defining the molecular structure that is already present within the data set.

10. A method according to any of the preceding claims wherein step (b) is performed using a method according to graph theory.

11. A method according to claim 10, wherein the method of performing step (b) in accordance with graph theory comprises:-

i) converting the first and second molecular structure data to first and second coloured graphs;

ii) determining a docking graph from the first and second graphs;

iii) identifying at least one clique within the docking graph; and

iv) converting each clique identified into molecular fragment data.

12. A method according to any of the preceding claims, further comprising ranking the molecular structure data in the data set according to the frequency with which molecular structures identical to each particular molecular structure have been

determined, and discarding the molecular structure data defining molecular structures which occur less frequently than a predetermined frequency threshold.

13. A method according to claim 13, wherein only a number of the molecular fragment data are retained within the database.

14. A method according to any of the preceding claims, further comprising:

f) comparing the molecular fragment data in the database with molecular structure data defining each of the predetermined molecular structures; and

g) determining the frequency of occurrence of each molecular fragment within each predetermined molecular structure.

15. A method according to claim 14, wherein the frequency includes fractional frequencies based upon the fraction of a particular molecular fragment present within a particular molecular structure.

16. A method according to claim 15, wherein the frequency of occurrence is determined for a particular predetermined molecular structure by:-

i) selecting the molecular structure data defining the particular predetermined molecular structure;

ii) selecting molecular fragment data representing a particular molecular fragment from the data set;

iii) comparing the selected molecular structure data and the molecular fragment data to determine common structure data representing the structure common to the predetermined molecular structure and the selected molecular fragment;

iv) determining the amount of the molecular fragment structure that the common structure data represents;

v) removing the determined common structure data from the predetermined molecular structure data; and

vi) repeatedly performing steps (iii) to (v) until no further common structure data is determined.

17. A method according to claim 16, wherein the comparison step (iii) is performed using a graph theory method.

18. A method of determining a relationship between the presence of a number of molecular fragments in a number of molecular structures and a biological target characteristic of the molecular structures, the method comprising:-

obtaining a modelling data set comprising data defining the molecular structures of a number of known molecules and corresponding known biological target characteristic data defining a common biological target characteristic for each molecule;

obtaining a database of molecular fragments data generated using a method according to any of the preceding claims;

obtaining data describing the presence of the molecular structures defined by the molecular fragment data, within the known molecules of the modelling data set; and,

determining a relationship between the data describing the presence of a number of the molecular fragments within the known molecules of the modelling data set and the common biological target characteristic data.

19. A method according to claim 18, wherein a number of the known molecules have identical structures to a number of the molecular structures used in the generation of the molecular fragment data.

20. A method according to claim 18 or claim 19, wherein the step of determining the relationship is performed using a numerical model.

21. A computer implemented method of generating predicted biological target characteristic data for a target molecule, the method comprising:-

obtaining target molecular structure data defining the molecular structure of the target molecule;

obtaining the relationship generated in the method according to any of claims 18 to 20;

processing the target molecular structure data to generate target fragment data describing the presence within the target molecule, of the molecular structures defined by the molecular fragment data used in the obtained relationship; and,

using the obtained relationship and the target fragment data to generate biological target characteristic data for the target molecule.

22. A method according to claim 21, wherein the molecular structure of the target molecule is different to the molecular structures of the known molecules or the molecules used in the generation of the molecular fragment data.

23. A method of determining a relationship between the presence of a number of molecular fragments in a number of molecular structures and a biological target characteristic, the method comprising:-

obtaining a modelling data set comprising data defining the molecular structures of a number of known molecules and corresponding known biological target characteristic data defining a common biological target characteristic for each molecule;

obtaining a database of molecular fragment data;

obtaining data describing the frequency of occurrence of a number of the molecular structures defined by the molecular fragment data, within the known molecules of the modelling data set wherein the data contains at least one non-integer frequency of occurrence; and

determining a relationship between the data describing the frequency of occurrence of the molecular fragments within the known molecules of the modelling data set, and the common biological target characteristic data.

24. A method according to claim 23, wherein at least the non-integer frequency of occurrence is determined for the molecular structure of a particular known molecule:-

i) selecting the molecular structure data defining the particular known molecule;

ii) selecting molecular fragment data representing a particular molecular fragment from the database;

iii) comparing the selected molecular structure data and the molecular fragment data to determine common structure data representing the structure common to the predetermined molecular structure and the selected molecular fragment;

iv) determining the amount of the molecular fragment structure that the common structure data represents;

v) removing the determined common structure data from the predetermined molecular structure data; and

vi) repeatedly performing steps (iii) to (v) until no further common structure data is determined.

25. A method according to claim 24, wherein the comparison step (iii) is performed using a graph theory method.

26. A method according to any of claims 23 to 25, wherein a number of the known molecules have identical structures to a number of the molecular structures used in the generation of the molecular fragment data.

27. A method according to any of claims 23 to 26, wherein the step of determining the relationship is performed using a numerical model.

28. A computer implemented method of generating biological target characteristic data for a target molecule, the method comprising:-

obtaining target molecular structure data defining the molecular structure of the target molecule;

obtaining the relationship generated in the method according to any of claims 23 to 27;

processing the target molecular structure data to generate target fragment data describing the presence within the target molecule, of the molecular structures defined by the molecular fragment data used in the obtained relationship wherein the presence includes at least one non-integer frequency of occurrence; and,

using the obtained relationship and the target fragment data to generate biological target characteristic data for the target molecule.

29. A method according to claim 28, wherein the molecular structure of the target molecule is different to the molecular structures of the known molecules or the molecules used in the generation of the molecular fragment data.

30. A computer program comprising program code means adapted to perform the method according to any of preceding claims, when the computer program is run on a computer.

31. A computer program according to claim 30, embodied on a computer readable medium.